KLASIFIKASI MULTI-LABEL PADA TOPIK AYAT AL-QURAN TERJEMAHAN BAHASA INGGRIS MENGGUNAKAN MULTINOMIAL NAIVE BAYES

A MULTI-LABEL CLASSIFICATION ON TOPICS OF QURANIC VERSES IN ENGLISH TRANSLATION USING MULTINOMIAL NAIVE BAYES

Reynaldi Ananda Pane [1], Mohamad Syahrul Mubarok [2], Adiwijaya[3]

[1,2,3] Prodi S1 Teknik Informatika, Fakultas Teknik, Universitas Telkom

[1]reynaldipane@students.telkomuniveristy.co.id,
[2]msyahrulmubarok@gmail.com,
[3]adiwijaya@telkomuniversity.ac.id

Abstrak - Al-Qur'an adalah firman Allah SWT yang menjadi kitab suci sekaligus pedoman bagi umat Islam di seluruh dunia yang setiap ayatnya mengandung makna dan hikmah. Ayat-ayat Al-Qur'an ada yang dapat diklasifikasikan ke dalam satu topik saja, namun ada juga yang dapat diklasifikasikan ke dalam beberapa topik yang berbeda. Hal ini termasuk ke dalam permasalahan klasifikasi multi-label. Tugas Akhir ini mengangkat permasalahan klasifikasi multi-label pada ayat-ayat Al-Qur'an tersebut menggunakan Multinomial Naïve Bayes sebagai classifier, serta dengan beberapa tahapan preprocessing data seperti case folding, tokenization, dan stemming. Hasil pengujian yang telah dilakukan menghasilkan nilai hamming loss terbaik sebesar 0.1247.

Kata Kunci: Al-Qur'an, klasifikasi *multi-label, multinomial naïve Bayes, hamming loss*

Abstract - Al-Qur'an is the words from Allah SWT which became the holy book as well as guidance for Muslims around the world. Each verse of the Qur'an contains meaning and wisdom that can usually be classified into more than one topic of discussion. In this Final Project will raise the issue of classification of Qur'anic topics that can be classified into more than one discussion as a multi-label classification problem, then this Final Project will produce a system that can receive input in the form of translations of the verses of the Qur'an in English and providing output in the form of topics classes contained in the verse. The system will be developed using the Multinomial Naïve Bayes classification method, with several stages of preprocessing data such as case folding, tokenization, and stemming. And also by using bag of words as feature extraction method. The best result from testing provide a 0.1247 hamming loss value.

Keywords: Al-Qur'an, multi-label classification, multinomial naïve bayes, hamming loss

1. Pendahuluan

Al-qur'an adalah kitab suci yang dijadikan pedoman serta petunjuk bagi umat Islam di seluruh dunia. Al-Qur'an terdiri dari 6236 ayat yang terbagi ke dalam 114 surah [1]. Setiap ayat dalam Al-Qur'an mengandung makna yang berbeda-beda, serta sering membahas lebih dari satu topik bahasan. Topik yang dibahas ayat-ayat Al-Qur'an dapat diklasifikasikan ke dalam 15 kelas [9], yaitu (1) Arkanul Islam, (2) Iman, (3) Al-Qur'an, (4) Ilmu dan Cabang-cabangnya, (5) Amal, (6) Dakwah, (7) Jihad, (8) Manusia dan Hubungan Kemasyarakatan, (9) Akhlak, (10) Peraturan yang Berhubungan dengan Harta, (11) Hal-hal yang Berkaitan dengan Hukum, (12) Negara dan Masyarakat, (13) Pertanian dan Perdagangan, (14) Sejarah dan Kisah-kisah, dan (15) Agama-agama.

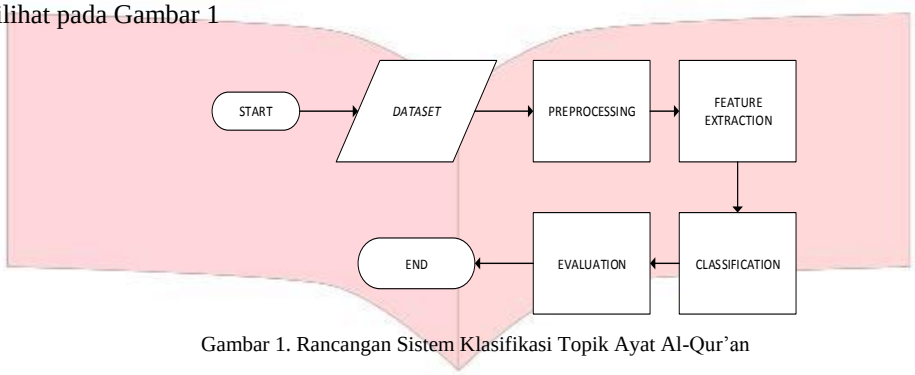
Hal menarik dari klasifikasi ayat Al-Qur'an adalah, setiap ayat dapat tergolong ke dalam lebih dari satu kelas yang berbeda. Fakta ini menunjukkan bahwa klasifikasi ayat Al-Qur'an berbeda dengan klasifikasi pada umumnya, dimana setiap data/dokumen hanya tergolong ke dalam sebuah kelas. Kasus klasifikasi seperti ini dapat disebut sebagai klasifikasi *multi-label*.

Penelitian terdahulu di bidang klasifikasi ayat Al-Qur'an cukup minim dan menggunakan *dataset* berupa huruf Arab asli [2]. Masih sangat sedikit ditemukan dataset dengan terjemahan ayat dalam bahasa Inggris. Pembangunan sistem dimulai dengan tahapan *preprocessing* yang dilanjutkan dengan ektraksi fitur, kemudian proses klasifikasi menggunakan *Multinomial Naive Bayes* yang disesuaikan untuk kasus klasifikasi *multi-label*, dan yang terakhir proses evaluasi. Metode *Multinomial Naive Bayes* dipilih karena sejauh ini teorema *Bayes* cukup baik digunakan untuk permasalahan klasifikasi teks [4] [5] [6] [7], akan tetapi model klasifikasi yang di bangun dengan teorema *Bayes* tersebut sejauh ini hanya digunakan untuk klasifikasi yang bersifat single-label.

Dataset yang digunakan adalah terjemahan ayat-ayat Al-Qur'an dalam bahasa Inggris. Pemilihan bahasa Inggris mengacu pada bahasa Internasional yang digunakan saat ini, sehingga hasil dari Tugas Akhir ini diharapkan dapat dimanfaatkan secara global bagi umat Islam di seluruh dunia

2. Perancangan Sistem

Perancangan sistem yang dibangun untuk klasifikasi *multi-label* pada topik terjemahan ayat Al-Quran dalam bahasa Inggris dapat dilihat pada Gambar 1



Gambar 1. Rancangan Sistem Klasifikasi Topik Ayat Al-Qur'an

Pembangunan sistem terbagi menjadi 3 tahap, yaitu : (1) *Preprocessing data*, (2) *Feature Extraction* dengan representasi *bag of words*, dan (3) *Classification* seperti pada Gambar 1.

a. *Preprocessing*

Tahapan ini bertujuan untuk membersihkan *dataset* dari *noise* agar siap digunakan dalam tahapan selanjutnya. Langkah awal adalah tahapan *case folding*. Dimana setiap huruf kapital yang ada di dataset dihilangkan sehingga setiap huruf di dalam dataset berubah menjadi huruf kecil *(lower case)*. Langkah berikutnya adalah *tokenization*, dimana langkah ini membagi teks pada kalimat-kalimat di *dataset* menjadi potongan kata tunggal [10]. Kemudian langkah berikutnya adalah *stopword removal*, dimana pada langkah ini kata-kata bantu seperti kata ganti orang, kata penghubung, dan kata yang tidak memiliki fungsi pada kalimat akan dihilangkan [15]. Langkah terakhir pada tahap *preprocessing* adalah *stemming*, tahapan ini berguna untuk mendapatkan bentuk dasar dari setiap kata dari beberapa kata yang memiliki bentuk dasar kata yang sama [13], yang didapatkan melalui proses sebelumnya.

b. *Feature Extraction*

Setelah mendapatkan *clean word* yang merupakan hasil dari tahapan *preprocessing*, maka langkah selanjutnya adalah melakukan ekstraksi fitur dengan representasi *bag of words*. Representasi ini dilakukan dengan cara menghitung jumlah kemunculan setiap kata berdasarkan kelas masing-masing yang terdapat pada data yang akan digunakan pada fase *training* [11].

c. *Classification*

Classifier dibangun dengan *Multinomial Naïve Bayes*. *Multinomial Naïve Bayes* merupakan salah satu jenis model teorema *Bayes* yang menggunakan distribusi multinomial [12]. Dalam menentukan sebuah kelas dari ayat, *Multinomial Naive Bayes* akan melakukan perhitungan probabilistik yang menentukan masuk atau tidaknya sebuah data ke dalam sebuah kelas tertentu. Perhitungan tersebut dapat dilihat pada persamaan 1.

$$\boldsymbol{P(c|d) \propto P(c) \prod_{1 \le k < n_d} P(t_k|c)} \quad (1)$$

dimana $P(c|d)$ adalah *posterior probability* dokumen *d* terhadap kelas *c*, $P(c)$ adalah *prior probability* kemunculan dokumen *c*, $P(t_k|\, c)$ adalah *conditional probability (likelihood)* dari munculnya *term (word)* t_k pada dokumen kelas *c*, dan n_d adalah jumlah fitur yang terdapat pada dokumen *d* [14].

d. *Evaluation*

Sebelumnya telah diuraikan bahwa kasus *multi-label* berbeda dengan kasus *single-label*. Dimana pada kasus *multi-label* setiap dokumen *d* akan dapat memiliki lebih dari satu kelas dari beberapa kelas yang terdapat di *dataset*. Untuk itu, dalam melakukan evaluasi terhadap hasil yang didapatkan diperlukan cara yang berbeda. Metode pengukuran *hamming loss* dipilih sebagai ukuran performa sistem yang dihasilkan. Nilai *hamming loss* dapat dihitung melalui persamaan 2.

$$Hamming\ Loss = \frac{1}{NL} \sum_{i=1}^{N} \sum_{j=1}^{L} [\hat{Y}_j^{(i)} \neq Y_j^{(i)}] \quad (2)$$

dimana N adalah jumlah data yang dianalisa, L adalah banyaknya label yang terdapat pada data yang dianalisa , $\hat{Y}_j^{(i)}$ adalah label ke-i milik data target ke- j, dan $Y_j^{(i)}$ adalah label ke-i milik data *output* ke-j

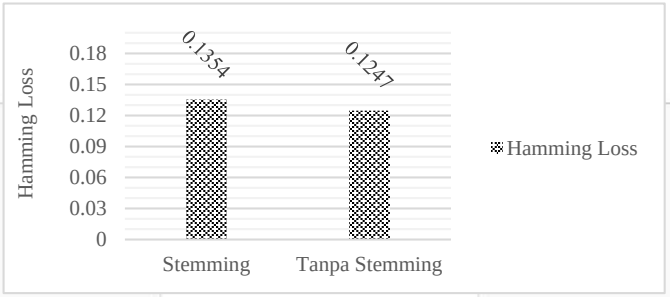
3. Hasil dan Pembahasan

Pada evaluasi terhadap performa *Multinomial Naive Bayes classifier*, dilakukan beberapa pengujian. Pengujian pertama berusaha mengetahui seberapa besar pengaruh metode *stemming* terhadap performa klasifikasi yang dihasilkan serta pengaruhnya terhadap waktu komputasi, Oleh karena itu, pada pengujian pertama dilakukan evaluasi terhadap preprocessing data yang dibedakan berdasarkan penggunaan metode *stemming* dan tanpa *stemming*. Data pada masing-masing jenis pengujian *preprocessing* dibagi ke dalam 5 *k-fold*. Pembagian data *training* dan *testing* pada setiap k-fold dapat dilihat pada Tabel 1.

Tabel 1. Jumlah data pada setiap k-fold

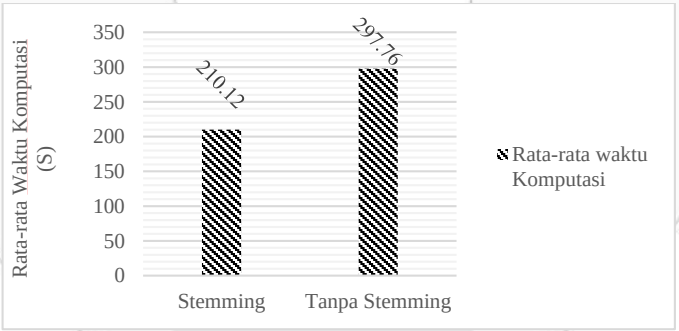
Fold	Data Train	Data Test
1	Ayat 1 – Ayat 4800	Ayat 4800 – Ayat 6236
2	Ayat 1200 – Ayat 6236	Ayat 1 – Ayat 1200
3	Ayat 1 – Ayat 1200 dan Ayat 2400 – Ayat 6236	Ayat 1200 – Ayat 2400
4	Ayat 1 – Ayat 2400 dan Ayat 3600 – Ayat 6236	Ayat 2400 – Ayat 3600

Hasil perbandingan nilai *hamming loss* terhadap masing-masing jenis *preprocessing* dapat dilihat pada Gambar 2.



Gambar 2. Perbandingan nilai hamming loss dengan menggunakan proses stemming dan tanpa stemming

Seperti yang terlihat pada Gambar. 2,dari hasil pengujian yang telah dilakukan. Didapatkan nilai *hamming loss* sebesar 0.1354 untuk klasifikasi yang melibatkan proses *stemming* pada tahapan *preprocessing*, sedangkan untuk klasifikasi yang tidak melibatkan proses *stemming* pada tahapan *preprocessing* didapatkan nilai *hamming loss* sebesar 0.1247. Sementara perbandingan waktu komputasi antar perbedaan penggunaan *stemming* dapat dilihat pada Gambar 3.

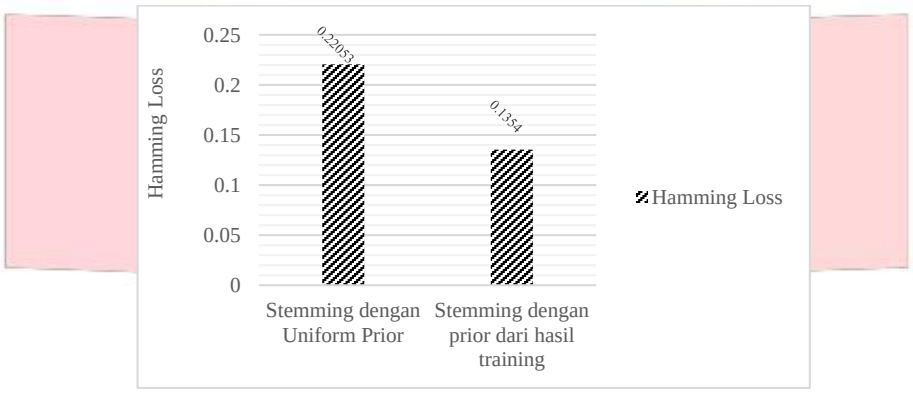


Gambar 3. Perbandingan rata-rata waktu komputasi dengan penggunaan stemming dan tanpa stemming

Berdasarkan hasil pada Gambar 3, terlihat bahwa rata-rata waktu komputasi dengan penggunaan *stemming* adalah 297.6 detik, sementara untuk rata-rata waktu komputasi tanpa penggunaan *stemming* yaitu 210.12 detik. Dengan begitu, proses *stemming* berhasil mempercepat rata-rata waktu komputasi dari sistem klasifikasi yang dibangun.

Dari hasil tersebut, maka proses klasifikasi yang tidak melibatkan proses *stemming* menghasilkan performa yang lebih baik. Hal ini menunjukkan bahwa kata-kata yang terkandung di Al-Qur'an memberikan ciri khusus terhadap setiap kelas, dengan kata lain memiliki pengaruh dalam menentukan kelas data. Sehingga generalisasi yang dilakukan berdasarkan bentuk dasar kata belum mampu memperbaiki performa *classifier*, karena menghilangkan kekhususan suatu kata pada suatu kelas. Akan tetapi, dalam hal waktu komputasi, proses *stemming* dapat menghemat waktu komputasi yang digunakan pada proses klasifikasi sebesar 29,44%. Hal ini dipengaruhi oleh metode *stemming* yang menghasilkan ekstraksi kata yang lebih sedikit.

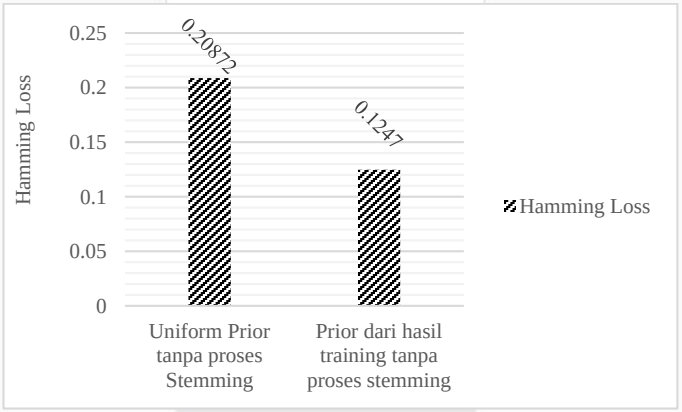
Selanjutnya, untuk menganalisa pengaruh nilai *prior probability* yang digunakan terhadap performa klasifikasi yang dihasilkan. Maka, pada pengujian kedua akan dilakukan evaluasi terhadap klasifikasi dengan menggunakan dua jenis nilai *prior probability* yang berbeda. Yaitu, nilai *prior probability* yang murni didapatkan dari hasil *training*. Dan jenis *prior* lainnya yang digunakan adalah *uniform prior probability*, nilai *prior* yang digunakan pada *uniform prior probability* adalah masing-masing 0.5 untuk $P(C = T)$ dan $P(C = F)$. Gambar 4 menunjukkan hasil perbandingan performa penggunaan kedua jenis *prior probability* ini, pada pengujian ini proses *stemming* dilibatkan pada tahap *preprocessing*.



Gambar 4. Perbandingan nilai hamming loss antara uniform prior dengan prior dari hasil training menggunakan stemming

Berdasarkan Gambar 4 , penggunaan nilai *prior* dari hasil *training* menghasilkan performa yang lebih baik. Terbukti dengan didapatkannya nilai *hamming loss* sebesar 0.1354 untuk klasifikasi yang melibatkan proses *stemming* pada tahapan *preprocessing* serta dengan menggunakan nilai *prior probability* yang didapatkan dari hasil *training*, sedangkan untuk klasifikasi yang menggunakan *uniform prior probability*, didapatkan nilai *hamming loss* sebesar 0.22053.

Berikutnya, diuji penggunaan kedua jenis *prior probability* ini tanpa melibatkan *stemming* pada fase *preprocessing*. Gambar 5 menunjukkan perbandingan nilai *hamming loss* yang dihasilkan dari penggunaan kedua jenis *prior*.



Gambar 5. Perbandingan nilai hamming loss antara uniform prior dengan prior dari hasil training tanpa melibatkan stemming

Dari hasil pengujian pada Gambar 5, terlihat bahwa penggunaan nilai *prior* yang didapatkan dari hasil *training* tetap menghasilkan performa klasifikasi yang lebih baik dibandingkan dengan menggunakan nilai *uniform prior*. Dari hasil yang diperoleh,proses klasifikasi yang menggunakan nilai *prior probability* dari hasil *training* menghasilkan performa yang lebih baik dari *uniform prior probability*. Ini berarti bahwa, keberagaman nilai *prior probability* yang didapat dari hasil *training* memiliki peranan yang cukup penting untuk menghasilkan *classifier* yang lebih baik, dibandingkan dengan menggunakan nilai *prior probability* yang seragam.

4. Kesimpulan

Penggunaan *stemming* dalam proses *preprocessing* belum mampu memperbaiki performa *classifier* dalam mengklasifikasikan data, meskipun perbedaan nilai *hamming loss* cukup kecil yakni sebesar 0.017. Hal ini menunjukkan bahwa kata-kata yang terkandung di Al-Qur'an memberikan ciri khusus terhadap setiap kelas, dengan kata lain memiliki pengaruh dalam menentukan kelas data. Sehingga generalisasi yang dilakukan berdasarkan bentuk dasar kata belum mampu memperbaiki performa classifier, karena menghilangkan kekhususan suatu kata pada suatu kelas. Tetapi penggunaan metode ini mampu mempercepat proses komputasi sebesar 29.44 %

Penggunaan nilai *prior probability* yang murni didapatkan dari hasil training menghasilkan performa klasifikasi yang lebih baik dari nilai *uniform probability*. Hal ini menunjukkan bahwa keberagaman nilai *prior* setiap kelas memegang yang cukup penting untuk performa *classifier*. Kombinasi nilai *prior probability* dari hasil *training* dengan tidak digunakannya stemming di fase *preprocessing* menghasilkan performa terbaik dengan nilai *hamming loss* 0.1247

Daftar Pustaka

[1] Syaamil Qur'an (2004). Cordova Al-Qur'an dan Terjemahan

[2] Al-Kabi, M. N., Ata, B. M. A., Wahsheh, H. A., & Alsmadi, I. M. (2013, December). A topical classification of Quranic Arabic text. In Proceedings of the Taibah University International Conference on Advances in Information Technology for the Holy Quran and its Sciences, Dec (pp. 22-25).

[3] Marcus, M.P., Marcinkiewicz, M.A. and Santorini, B., 1993. Building a large annotated corpus of English: The Penn Treebank. Computational linguistics, 19(2), pp.313-330.

[4] Mubarok, M. S., Adiwijaya, & Aldhi, M. D. (2017, August). Aspect-based sentiment analysis to review products using Naïve Bayes. In AIP Conference Proceedings (Vol. 1867, No. 1, p. 020060). AIP Publishing.

[5] Aziz, R. A., Mubarok, M. S., & Adiwijaya, A. (2016, September). Klasifikasi Topik pada Lirik Lagu dengan Metode Multinomial Naive Bayes. In Indonesia Symposium on Computing (IndoSC) 2016.

[6] Saputra, A., Adiwijaya, A., & Mubarok, M. (2017). Klasifikasi Sentimen Pada Level Aspek Terhadap Ulasan Produk Berbahasa Inggris Menggunakan Bayesian Network (case Study: Data Ulasan Produk Amazon). eProceedings of Engineering, 4(3)

[7] Syahnur, M. H., Bijaksana, M. A., & Mubarok, M. S. (2016). Kategorisasi Topik Tweet Di Kota Jakarta, Bandung, Dan Makassar Dengan Metode Multinomial Naïve Bayes Classifier. eProceedings of Engineering, 3(2).

[8] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

[9] M.H. Shakir. Al-Qur'an English Translation. [Online] Available at : http://www.theholyquran.org/?x=s_main&kid=15

[10] Korenius, T., Laurikkala, J., Järvelin, K. and Juhola, M., 2004, November. Stemming and lemmatization in the clustering of finnish text documents. In Proceedings of the thirteenth ACM international conference on Information and knowledge management (pp. 625-633). ACM.

[11] Putri, L., Mubarok, M., & Adiwijaya, A. (2017). Klasifikasi Sentimen Pada Ulasan Buku Berbahasa Inggris Menggunakan Information Gain Dan Naïve Bayes. eProceedings of Engineering, 4(3).

[12] Mubarok, M. S., & Asriadie, M. S. Klasifikasi Emosi Pada Twitter Menggunakan Bayesian Network.

[13] Julianto, B., Adiwijaya, A., & Mubarok, M. (2017). Identifikasi Parafrasa Bahasa Indonesia Menggunakan Naive Bayes. eProceedings of Engineering, 4(3).

[14] Prayuga, N., Adiwijaya, A., & Mubarok, M. (2017). Klasifikasi Polycystic Ovary Syndrome Berdasarkan Citra Ultrasonografi Menggunakan Principal Component Analysis Dan Naïve Bayes Untuk Membantu Mendeteksi Kesuburan Wanita. eProceedings of Engineering, 4(3).

[15] Sitompul, D., Adiwijaya, A., & Mubarok, M. (2017). Analisis Sentimen Level Kalimat Pada Ulasan Produk Menggunakan Bayesian Networks. eProceedings of Engineering, 4(3).